Statistical validity of water quality time series in urban watersheds

Validade estatística de séries históricas de qualidade da água em bacias urbanas

Marcelo Coelho[1], Cristovão Vicente Scapulatempo Fernandes[1], Daniel Henrique Marco Detzel[1] and Michael Mannich[1]

[1]Universidade Federal do Paraná, Curitiba, PR, Brazil

E-mails: mcoelho2011@hotmail.com (MC), cris.dhs@ufpr.br (CVSF), daniel@lactec.org.br (DHMD), mmannich@gmail.com (MM)

ABSTRACT

The water resources quality continuous monitoring is a complex activity. It generates extensive databases with time series of many variables and monitoring points that require the application of statistical methods for the information extraction. The application of statistical methods for frequency analysis of time series is linked to attending of the basic assumptions of randomness, homogeneity, independence, and stationarity. However, despite its importance, the verification of these assumptions in water quality literature is unusual. Therefore, the present study tests the Upper Iguaçu basin water quality time series against the mentioned hypotheses. Rejection was observed in 15%, 26%, 51% e 31% for randomness, homogeneity, independence, and stationarity, respectively. The results evidenced the strong relation between monitoring strategy, data assessment and meeting of basic statistical assumptions for the analysis of water quality time series. Even with the existence of possible solutions for addressing those issues, the standard monitoring strategies, with irregular frequencies and lack of representativeness in relation to other periods, beyond commercial, act as an obstacle to their implementation.

Keywords: Homogeneity; Independence; Randomness; Stationarity; Monitoring strategies.

RESUMO

O monitoramento quali-quantitativo de recursos hídricos é uma atividade complexa. Sua continuidade dá origem a extensos bancos de dados contendo séries históricas de diversas variáveis e pontos, que exigem a aplicação de métodos estatísticos para extração de informações. A aplicação de métodos estatísticos para a análise de frequências de séries históricas está vinculada ao atendimento aos pressupostos básicos de aleatoriedade, homogeneidade, independência e estacionariedade. Apesar da importância destes critérios, sua verificação é incomum em literatura de qualidade de água. Neste artigo, estas hipóteses foram testadas em séries históricas da Bacia do Alto Iguaçu. Observou-se rejeição em 15%, 26%, 51% e 31% das séries testadas para aleatoriedade, homogeneidade, independência e estacionariedade, respectivamente. Os resultados evidenciaram a forte relação entre as estratégias de monitoramento, a forma de abordagem dos dados e o atendimento aos pressupostos básicos para análise de séries de qualidade da água. Apesar da existência de possibilidades para solução do problema de não atendimento, as estratégias típicas de monitoramento com frequências irregulares e pouca representatividade em relação a outros períodos, além do horário comercial, representam um obstáculo para sua implementação.

Palavras-chave: Homogeneidade; Independência; Aleatoriedade; Estacionariedade; Estratégias de monitoramento.

INTRODUCTION

Qualitative monitoring of water resources is a complex activity. It requires the development, constant maintenance and updating of strategies for obtaining representative data and consistent information for the management of water resources. Continuity of monitoring gives rise to extensive databases containing time series of various variables and monitoring points. The analysis of these data requires the application of appropriate statistical methods, especially in urban basins, where processes that alter water quality are diverse and produce high variability.

The application of statistical methods for the analysis of frequencies in time series depends fundamentally on the statistical validity of the series, which is linked to the basic assumptions of randomness, homogeneity, independence and stationarity (CHADEE; SHARMA, 2003; MERZ; THIEKEN, 2005; NAGHETTINI; PINTO, 2007; HULLEY; CLARKE; WATT, 2015). In hydrological studies, the verification of these hypotheses is frequently recommended, however, in water quality literature references that have some recommendation or verification are rare.

Water quality time series are used in trend analysis (COSTA; OLIVEIRA; SOUZA, 2011; DAHM et al., 2014), multivariate statistical analyses (WANG et al., 2013; TORRES; LEMOS; MAGALHÃES JUNIOR, 2016), analysis of variances (GONÇALVES et al., 2011), mathematical modeling (KONDAGESKI; FERNANDES, 2009; KNAPIK et al., 2013; VIEIRA et al., 2013), duration curves (MARIN et al., 2007; FORMIGONI et al., 2011), linear regression analysis (HIRSCH; MOYER; ARCHFIELD, 2010), among other analyses, however, typically, without prior verification of the assumptions for analysis.

Costa, Oliveira and Souza (2011) analyzed the presence of trends in time series of annual averages of water quality data in the Upper Iguaçu Basin. The annual averages were used as a solution to the problem of dependence and seasonality in the series. However, no verification of compliance with the basic assumptions for frequency analysis in time series was performed, and therefore there is no guarantee that the annual averages are statistically representative.

The concepts regarding the assumptions are clearly defined in hydrological studies. Randomness is observed when fluctuations of values have natural causes, such as river flows when not influenced by the presence of a dam. Independence is observed when previous observations do not influence the next observations in a chronological sequence. A homogeneous time series is observed when all the data comes from a single population. Different populations can be generated in the same series by the influence of phenomena such as *El Niño* or seasonality. Finally, a series is said to be stationary when there is no changing of population parameters, such as mean and standard deviation, over time (NAGHETTINI; PINTO, 2007).

In water quality studies there is no clear and appropriate definition of these concepts to explain cause and effect relationships. This absence is equivalent to considering the series random, independent, homogeneous and stationary, but without proper justification. Environmental data are susceptible to the presence of non-randomness, dependence, non-homogeneity, and non-stationarity, due to seasonality, the occurrence of extreme events, urbanization, among other factors. Hence, the need for a more cautious analysis, especially in urban basins, is evident.

The negligent application of traditional statistical methods in series with these characteristics produces unreliable information on parameters and their associated uncertainties, that is, on confidence intervals, p-values, standard deviation, among others (GILBERT, 1987; HELSEL; HIRSCH, 2002; MCBRIDE, 2005).

In this paper, a verification of the basic statistical assumptions for frequency analysis in water quality time series of the Upper Iguaçu Basin is performed. These series represent the technical basis of several relevant studies for the management of water resources. The results are accompanied by a critical analysis of the application to water quality data, the discussion of causes and effects of non-attendance, and recommendations on the development of monitoring strategies and approaches to the observed series.

MATERIAL AND METHODS

The research was developed from a database consolidated between 2005 and 2011 for research projects in the Upper Iguaçu Basin and complemented with five monitoring campaigns in 2012. The time series were submitted to verification of the attendance to the assumptions for the analysis of frequencies by the application of statistical methods, testing the hypotheses of randomness, homogeneity, independence, and stationarity.

Prior to the verification of compliance with the basic assumptions, characteristics of the time series, such as sample size, variability, sampling frequencies and seasonality, were analyzed.

Study area

The Upper Iguaçu Basin (Figure 1) is located in the State of Paraná, in the Metropolitan Region of Curitiba (RMC). It covers 15 municipalities, among which, Curitiba, the capital of the State of Paraná. Its drainage area is approximately 3,000 km² and is located between latitudes -25.23 and -25.83 and longitudes -48.96 and -49.69. The Iguaçu River crosses the State of Paraná in the east-west direction, extending for 1,320 km, of which the initial 90 km are in the Upper Iguaçu Basin.

Urbanization is mainly concentrated in the Barigui, Belém, Atuba, and Palmital rivers, on the right bank of the river. Approximately 15% of the area of the basin is urbanized. The other uses of the soil in the basin are mainly divided between temporary crops, fields and other types of vegetation, with the predominance of agricultural uses, thus leaving little forest cover on the basin (COELHO, 2013).

Research and database

The detailed analysis of water quality time series, involving questioning its representativeness and validity for subsequent frequency analysis, is original and distinct from other studies in the Upper Iguaçu Basin. Studies such as Kondageski and Fernandes (2009), Pitrat (2010), Knapik et al. (2013), and Costa; Oliveira and Souza (2011), used water quality time series in multivariate statistical analysis, trend analysis, mathematical modeling of water
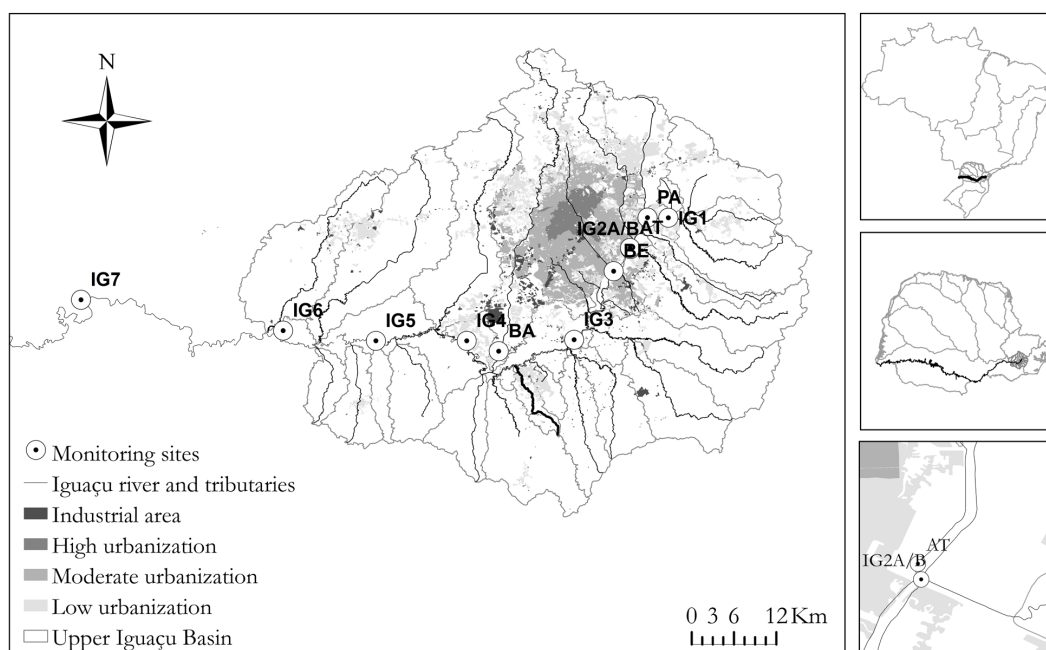
Figure 1. Upper Iguaçu Basin and monitoring sites.

Table 1. Monitoring sites information.

Site	River	Latitude	Longitude	Area**
IG1	Iguaçu	-25.4433	-49.1406	283
IG2A/B*	Iguaçu	-25.4833	-49.1892	626
IG3	Iguaçu	-25.5989	-49.2608	1,284
IG4	Iguaçu	-25.6003	-49.3978	2,122
IG5	Iguaçu	-25.6003	-49.5133	2,578
IG6	Iguaçu	-25.5872	-49.6317	3,049
IG7	Iguaçu	-25.5481	-49.8894	3,662
PA	Palmital	-25.4413	-49.1685	102
AT	Atuba	-25.4738	-49.1855	142
BE	Belém	-25.5074	-49.2151	94
BA	Barigui	-25.6136	-49.3566	257

* IG2A left river bank, IG2B right river bank; ** Unit in km².

quality, generation of duration curves, box plots, among other analyses, however, without worrying about the issues discussed in this article.

These surveys contributed to the consolidation of a database currently composed of 12 monitoring sites (Table 1), with 8 sites in the Iguaçu River and 4 in the main tributaries, presented in Figure 1, with 34 water quality variables.

Monitoring

The monitoring sites coincide with sites in the national information system on water resources, which are officially monitored by the local water agency, the Instituto das Águas do Paraná. It is important to emphasize that the monitoring described here is independent of the monitoring performed by that institution, therefore, not necessarily following the same methods and techniques.

The analyzed database is composed of data between 2005 and 2012 (43 monitoring campaigns), except in the tributaries (PA, AT, BE and BA, see Table 1) of the Iguaçu river and IG7, monitored only in 2012 by the present research, between August and December, in a monthly time step (5 campaigns).

The monitoring campaigns were typically held on business days between 8:00 am and 5:00 p.m. There were no records of campaigns during the night, weekends and holidays.

Flow estimates

All points, except IG1, AT and BE, have a fluviometric station, whose data and time series obtained by the Instituto das Águas do Paraná, are available in the Brazilian national water information system (Hidroweb). They also have a time series measured by this and other previously cited studies, which is included in the present work.

In years prior to 2012, the flows were estimated from a single reading of the ruler and later entered into a rating curve provided by the Instituto das Águas do Paraná. However, in 2012, it was found that most stations presented problems such as lack of ruler, large cross-sectional changes, and outdated rating curves. Thus, for the determination of flow rates in 2012, it was used the average specific flow rate for the Upper Iguaçu Basin (Equation 1) presented in Marin et al. (2007).

$$q = -0,0178 \ln p + 0,084 \qquad (1)$$

q = average specific flow rate (m³/s.km²);
p = permanence (%).

The specific flow q was calculated as a function of the continuous flow of a reference site. Due to better conditions (rules, location, and recommendations of network users) IG7 (code Hidroweb 65035000 - Porto Amazonas) was used as a

reference. From its flow time series (1970-2010), a duration curve was determined. With this, the IG7 flow permanence values in each campaign (calculated by the rating curve) were obtained. Then, q was calculated and multiplied by the drainage area of each site.

Water quality analysis

All laboratory analyzes followed the procedures described in APHA (1999). The analyzed variables were: Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Dissolved Organic Carbon (COD), Dissolved Oxygen (OD and ODWINKLER), Total Solids (STT), Fixed Total Solids (STF), Volatile Total Solids (STV), Total Dissolved Solids (SDT), Fixed Dissolved Solids (SDF), Volatile Dissolved Solids (SDV), Total Suspended Solids (SST), Fixed Suspended Solids (SSF), Volatile Suspended Solids (SSV), Settleable Solids (S_{SED}), Ammoniacal Nitrogen (NH_4^+), Organic Nitrogen (N_{ORG}), Kejhdahl Nitrogen ($N_{KEJLDAHL}$), Nitrite (NO_2^-), Nitrate (NO_3^-), Total Nitrogen (N_T), Total Phosphorous (P_T), Ortophosphate (PO_4^{3-}), Dissolved Organic Phosphorous ($P_{ORG\ DISS}$), Dissolved Phosphorous (P_{DISS}), Particulate Phosphorous (P_{PART}), Total Coliforms ($COLIF_{TOTAL}$), Fecal coliforms ($COLIF_{FECAL}$). The variables Conductivity (COND), Turbidity (TURB), Hydrogen potential (pH), Temperature (T) and Secchi Disk Depth (SECCHI) were measured in field by sensors for COND, TURB, T and pH, calibrated at each collection, and Secchi disk for SECCHI. More detailed information on methods, limits of detection and brand of sensors can be found in Coelho (2013).

Hypothesis testing

The hypotheses of randomness, homogeneity, independence, and stationarity were tested by the following methods: Runs test, Wilcoxon-Mann-Whitney, Wald and Wolfowitz and Kendall tests, respectively. The presence of seasonality was evaluated by Kruskal-Wallis and median tests. The tests were applied in the time series of 34 monitoring variables in terms of quantity (loads) and quality (concentrations and other units), for 12 monitoring sites, with significance level α = 5%. The Wilcoxon-Mann-Whitney, Kruskal-Wallis, and median series tests were performed in SPSS Statistics 20 program from IBM®; the Wald and Wolfowitz test in Excel®; and the Kendall in XLSTAT 2013 program of Addinsoft®.

Runs test

The null hypothesis (H_0) that the series is random against the alternative hypothesis (H_1) that the series is not random was tested. Let *m* be the number of elements of a type and *n* the number of elements of another type in a sequence of $N = m + n$ binary events. The test consists of counting the number of groups of equal elements in a sequence, or number of runs (*r*), for example, $+ + - - -$ ($r = 2$, $m = 2$, $n = 3$), and comparison with expected values in function of *m* e *n* (SIEGEL; CASTELLAN JUNIOR, 2006).

In order to transform the numerical historical series *X* into binary events, the median was used as the parameter, that is, if $x_i >$ median, $N_i = +$, if $x_i <$ median, $N_i = -$. If both *m* and *n* are ≤ 20, a specific table is used, with the expected minimum and maximum values of r as a function of *m* and *n*. If *m* or *n* is > 20, the normal distribution is used as an approximation of the distribution of *r* by Equations 2, 3 and 4. The test is bilateral and H0 is rejected when p-value <0.05 (SIEGEL; CASTELLAN JUNIOR, 2006).

$$\mu_r = \frac{2mn}{N} + 1 \tag{2}$$

$$\sigma_r = \sqrt{\frac{2mn(2mn - N)}{N^2(N - 1)}} \tag{3}$$

$$z = \frac{r - \mu_r}{\sigma_r} \tag{4}$$

Wilcoxon-Mann-Whitney

In this test, the null hypothesis (H_0) is that the series is homogeneous, while the alternative hypothesis (H_1) is that the series is not homogeneous.

Let *m* be the number of cases in the sample of the group X and *n* the number of cases in the sample of the group Y. Assuming that the samples are independent, the observations of both groups are combined by arranging the stations in ascending order. The value of the test statistic, W_x, is the sum of the posts of the smallest group (SIEGEL; CASTELLAN JUNIOR, 2006).

When *m* and *n* are ≤ 10, a specific table is used to determine the exact probability associated with occurrence of any W_x as extreme as the observed, when H_0 (the series is homogeneous) is true. For *m* or *n* > 10 the distribution of W_x approaches a normal distribution and the significance of an observed value of W_x can be determined by Equations 5, 6 and 7 (SIEGEL; CASTELLAN JUNIOR, 2006).

$$\mu_{W_x} = \frac{m(N + 1)}{2} \tag{5}$$

$$\sigma_{W_x}^2 = \frac{mn(N + 1)}{12} \tag{6}$$

$$z = \frac{W_x \pm 0,5 - \mu_{W_x}}{\sigma_{W_x}} \tag{7}$$

In the context of water quality time series, the groups X and Y are the first and second half of the series, respectively. The test is bilateral and H_0 is rejected when p-value < 0.05.

Wald and Wolfowitz

Considering a sample $\{X_1, X_2, ..., X_N\}$ and the differences $\{X_1', X_2', ..., X_N'\}$ determined between X_i and its average $\bar{X}$, the Equation 8 is calculated. Under an independence hypothesis, *R* follows a normal distribution with mean and variance given by Equations 9 e 10, respectively. The test statistic is then determined by Equation 11. H_0 is rejected if $|T| > z_{1-\alpha/2}$ (NAGHETTINI; PINTO, 2007).

$$R = \sum_{i=1}^{N-1} X'_i X'_{i+1} + X'_1 X'_N \tag{8}$$

$$E[R] = -\frac{s_2}{N-1} \tag{9}$$

$$VAR[R] = \frac{s_2^2 - s_4}{N-1} + \frac{s_2^2 - 2s_4}{(N-1)(N-2)} - \frac{s_2^2}{(N-1)^2} \tag{10}$$

$$s = N m_r'$$

$$m_r' = \frac{\sum_{i=1}^{N} \left(X'_i\right)^r}{N}$$

$$T = \frac{R - E[R]}{\sqrt{VAR[R]}} \tag{11}$$

Seasonal Kendall

Let x_{il} be the measures of station i of year l, K the number of stations and L the number of years. The seasonal Kendall tests for a monotonic trend in the data, considering different seasons. Hence, distinct trends may be identified for distinct seasons.

Considering the seasons as the months of the year, for each month the observations of each year are used to calculate the statistic S_i and S statistic according to Equations 12 and 13, respectively. The variance of S is calculated by Equation 14, where n_i is the number of observations in station i. After S and $VAR(S)$ are calculated, Z is calculated according to Equation 15. In the bilateral case H_0 is rejected if $|Z| > Z_{1-\alpha/2}$. In the unilateral case H_0 is rejected in favor of the presence of an increasing trend if $Z > 0$ and $|Z| > Z_{1-\alpha}$, or in favor of a decreasing trend if $Z < 0$ and $|Z| > Z_{1-\alpha}$, with Z following a normal distribution (GILBERT, 1987; HELSEL; HIRSCH, 2002). Due to the presence of dependence in the series, the Hamed and Rao (1998) correction was applied by the software.

$$S_i = \sum_{k=1}^{n_i-1} \sum_{l=k+1}^{n_i} sgn(x_{il} - x_{ik}) \tag{12}$$

$$sng(x_{il} - x_{ik}) = 1 \ if \ x_{il} - x_{ik} > 0$$

$$= 0 \ if \ x_{il} - x_{ik} = 0$$

$$= -1 \ if \ x_{il} - x_{ik} < 0$$

$$S = \sum_{i=1}^{K} S_i \tag{13}$$

$$VAR(S) = \sum_{i=1}^{K} \frac{1}{18}\left[n_i(n_i-1)(2n_i+5)\right] \tag{14}$$

$$Z = \frac{(S'-1)}{\left[VAR(S')\right]^{1/2}} \ if \ S' > 0$$

$$= 0 \ if \ S' = 0$$

$$= \frac{(S'+1)}{\left[VAR(S')\right]^{1/2}} \ if \ S' < 0 \tag{15}$$

Kruskal-Wallis

The Kruskal-Wallis tests the null hypothesis that the median of k data groups are the same, against the alternative hypothesis that at least one pair of groups presents differences between their medians. Each of the N observations of the k groups is replaced by ranks, i.e. the rank 1 is assigned to the lowest value and the rank N to the highest, considering observations as a unique series. Then, the ranks average is calculated in each group and in all groups combined.

The test statistic is given by Equation 17, where N is the number of observations in the combined sample, k is the number of groups, n_js the number of cases in the j-th group, $\overline{R_j}$ is the average of ranks in the j-th group and $\overline{R} = (N+1)/2$. For $k > 3$ and $n_j > 5$ in each group, the sample distribution of KW is approximated by the χ^2 distribution with $k-1$ degrees of freedom. H_0 is rejected if $KW \geq$ to the tabulated value of χ^2 (SIEGEL; CASTELLAN JUNIOR, 2006).

$$KW = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left(\overline{R_j} - \overline{R}\right)^2 \tag{17}$$

Medians test

Finally, the last considered method tests similar hypotheses than the Kruskal-Wallis test. The median of the combined sample of the k groups is determined and the observations higher and lower than the median are replaced by + and −, respectively. The test statistic is given by Equation 18, where $i = 1$ for + category, or 2 for −., category, n_{ij} the number of cases in the each group categories j, k the number of groups, and E_{ij} is the avegare between n_{1j} and n_{2j}. The obtained values $\times^2$, for large samples, are distributed as a χ^2 with $(r-1)(k-1)$ degrees of freedom, where r is the number of categories and k the number of groups (SIEGEL; CASTELLAN JUNIOR, 2006).

$$\times^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{\left(n_{ij} - E_{ij}\right)^2}{E_{ij}} \tag{18}$$

For the application of the Kruskal-Wallis and the medians tests, four groups were defined, corresponding to the four seasons of the year.

RESULTS E DISCUSSION

Time series characteristics

Figure 2 shows a small amount of summer data, when compared to the other seasons. In the same sense, fewer data from tributaries is considered in comparison with the main river data. The time series presented irregular frequency between 2005 and 2013 (Figure 3), with intervals from biweekly to quarterly. Long

intervals without monitoring are observed between 2006-2008, 2008-2009 and 2011-2012.

The phosphorus series, except P_T, as well as $COLIF_{TOTAL}$, $COLIF_{FECAL}$, COD, N_T, $OD_{WINKLER}$ and S_{SED} series, have a considerably smaller amount of data in relation to other variables at all sites. In terms of loads, some series are smaller because of the lack of flow data, specially at sites IG1, IG3, IG4 e IG6.

The boxplots for concentration, load, and other variables (T, pH, COND, TURB, Q, S_{SED}, $COLIF_{TOTAL}$, $COLIF_{FECAL}$ e SECCHI), in general, showed high variability, a large amount of outliers, and positive asymmetry, suggesting non-normality of the series. Some of these characteristics can be observed in the flow boxplots (data between 2005-2012) presented in Figure 4. The basin presents a considerable increase of loads and flows downstream the site IG3.
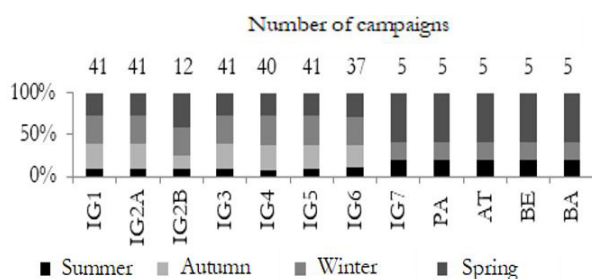

Figure 2. Seasonal distribution of campaigns.

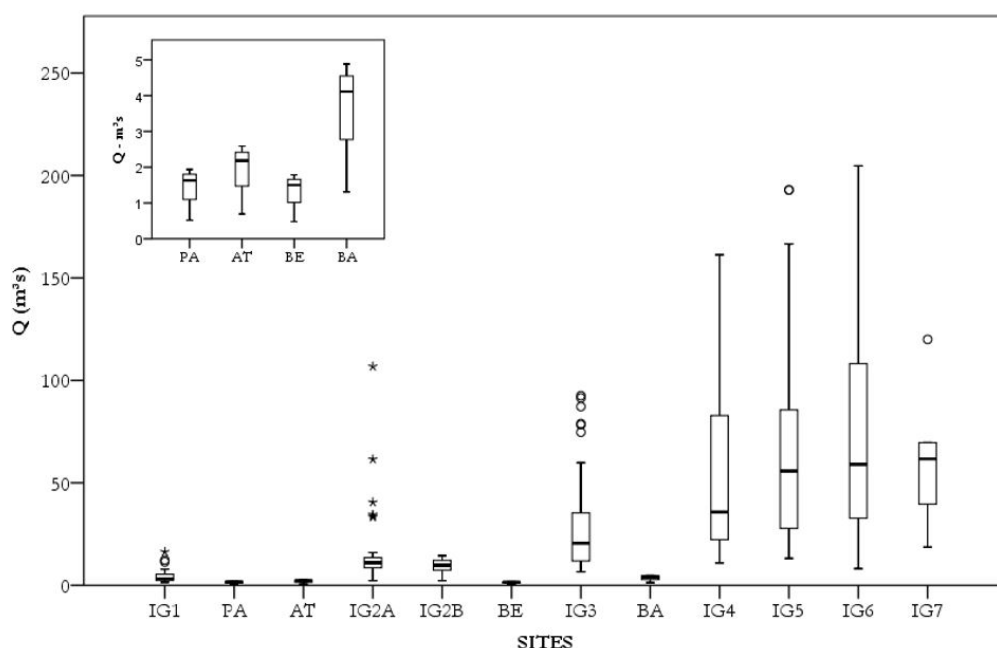**Figure 3.** Distribution of sampling frequency (each line is a campaign).

Hypothesis testing

The results of the hypothesis tests are presented in this section. Except in Table 2, in each parameter, the 1st row contains the results for the water quality time series and the 2nd row, the results for the quantity time series. Variables absent from the tables showed no rejection at any monitoring site, and missing points did not show rejection for any variables.

The tests were applied in a total of 720 series (34 variables x 12 sites for quality and 26 variables x 12 sites for loads). The analysis of the results is presented in terms of percentage of rejections and p-values, however, the sites IG7, PA, AT, BE and BA did not present rejections in the tests of seasonality, randomness, and homogeneity, possibly due to the low amount of data (5 observations) and, consequently, lack of evidence for rejection. For these tests the total of series considered for calculation of the percentages of rejection is 420.

Seasonality

Table 2 shows that the seasonal cycle exerts a significant influence on 9% of the series. In most cases, rejections are observed in both tests. The rejections in NH_4^+ and N_{ORG} series from IG3, T in all sites, and some forms of solids in IG2A are highlighted.

According to the results, the seasonal cycle seems to exert a greater influence on the organic contribution, represented by variables such as SSV, NH_4^+, N_{ORG} and $N_{KEJLDAHL}$, than in the hydrological cycle (Q). The smaller influence on the hydrological cycle is associated with the form of operation to collect information, performed only on business days without the concern of collecting during the occurrence of rainfall. On the other hand, the greater influence on water quality is associated to significant temperature

Figure 4. *Boxplots* of flow rates.

Table 2. Rejection of H_0 (equality between seasons) in seasonality tests.

Variables	IG1	IG2A	IG2B	IG3	IG4	IG5	IG6
DQO	M2	*	*	*	M2	*	*
DBO	*	*	*	K2	*	K1M1	*
STT	*	**K1M1**	*	*	*	*	*
STF	*	**K1**	*	*	*	*	*
SDT	*	**K1M1**	*	*	*	*	*
SDF	*	**K1**	*	*	*	*	*
SSF	*	**M2**	*	*	*	*	*
SSV	*	*	*	*	K2M2	*	K2
NH_4^+	*	K1	*	**K1M1**	**K1M1**	**K1M1**	**K1M1**
N_{ORG}	K2M2	*	*	*	*	**K1M1**	M2
$N_{KEJLDAHL}$	M2	*	*	**K1**	**K1M1**	**K1M1**	**K1M2**
NO_2^-	*	K1	*	*	*	*	M1
N_T	*	*	*	*	*	*	M2
TURB	K	*	*	M	*	*	*
OD	*	M1	*	*	*	*	*
T	**KM**	**KM**	**K**	**KM**	**KM**	**KM**	**KM**
Q	KM	*	*	*	*	*	*

K: Kruskal-Wallis; M: Median test; 1: water quality series; 2: water quantity series. Example: K1M1: rejection of H_0 in both tests in water quality series. Some variables do not present numerical distinction, since the load calculation does not apply.
* Non-rejection of H_0.

Table 3. Rejections of H_0 (series is random) in randomnsess test.

Variables	IG1	IG2A	IG2B	IG3	IG4	IG5	IG6
DQO	0.037	*	*	0.037	*	*	*
	0.000	*	*	*	*	*	*
DBO	0.004	*	*	0.023	*	0.023	0.002
	0.005	*	*	*	*	0.033	*
STV	*	0.011	*	*	*	*	*
	*	*	*	*	*	*	*
SST	*	0.004	*	0.004	*	*	*
	*	0.005	*	*	*	*	*
S_{SED}	*	0.010	*	*	*	*	*
	*	0.028	*	*	*	*	*
NH_4^+	*	*	*	0.037	*	*	0.006
	*	0.028	*	*	*	*	*
N_{ORG}	*	*	*	*	*	0.023	*
	*	0.046	*	*	*	0.033	*
$N_{KEJLDAHL}$	*	*	*	*	*	0.004	*
	*	*	*	*	*	*	0.005
NO_2^-	0.000	0.000	*	*	0.037	*	*
	0.005	0.005	*	0.040	*	*	*
N_T	*	*	*	*	0.046	*	*
	*	*	*	*	*	*	*
P_T	0.009	0.011	*	*	0.006	*	0.020
	*	*	*	*	*	*	*
P_{DISS}	*	*	0.006	*	*	*	*
	*	*	*	*	*	*	*
P_{PART}	*	*	*	*	*	*	0.033
	*	*	*	*	*	*	*
$P_{ORG\ DISS}$	*	*	*	*	*	*	*
	*	*	*	*	0.011	*	*
COD	*	*	*	*	*	*	*
	*	0.026	0.006	*	*	*	*
TURB	*	0.037	*	0.001	*	*	*
COND	**0.000**	**0.000**	*	**0.000**	**0.000**	**0.000**	**0.000**
OD	*	0.046	*	0.046	0.029	*	*
	0.012	*	*	*	*	*	*
T	0.028	0.027	*	0.004	0.006	0.037	0.003
pH	0.000	0.011	0.034	*	*	0.023	0.046
Q	0.000	0.000	*	0.040	*	*	*

* Non-rejection of H_0.

variations and to diffuse pollution, since in the rainy periods, mainly with the first rains, the contribution of diffuse pollution increases.

Therefore, interferences in the physical, chemical and biological processes throughout the season are expected. This emphasizes the importance of the representativeness of the series, that is, the series should represent all the variability of the parameter or variable being analyzed. The representativity is associated to the frequency of information collections. Thus, irregular collections of flow data will hardly represent their temporal variability accurately.

The presence of trend in the series introduces high variability in the groups referring to each station when older data are grouped to the most recent. This makes it difficult to differentiate the groups.

Randomness

It is observed in Table 3 that the randomness hypothesis was rejected in 15% of the series, half with p-value <0.01, especially the COND variable. The low p-values for COND can be explained by the apparent trend of the series, shown in Figure 5.

It is observed that from 2008 the values of conductivity are higher than the median, unlike the older ones. Since the method consists in analyzing the number of oscillations (runs) around the median (4 in this case), the rejections are due to the low number of runs.

This result shows that non-randomness may be a consequence of the analysis of long periods in which the series are non-stationary or non-homogeneous. The increasing trend and non-homogeneity in these series are confirmed by trend and homogeneity tests applied.

The causes for these effects may be real changes in water quality conditions or changes in the form and/or frequency of measurements, methods of analysis, etc., i.e. changes in monitoring strategy. It emphasizes the importance of investigating the origin of non-randomness.

Homogeneity

Table 4 shows the rejection of homogeneity in 26% of the series, half with p-value <0.01. Once again the COND variable is highlighted. Even if the chance of type I error (rejecting H_0 when this is true) was 10% (α), there would be rejection in several series.

The test application method consists of dividing the series in half and comparing the two groups, so the rejections indicate high temporal variability of the series and possible non-stationarity (trend).

The non-homogeneity implies the presence of two or more populations in the same sample, so the p-values, confidence intervals, and other statistical inferences obtained by the analysis of this type of time series do not represent any of the tested populations.
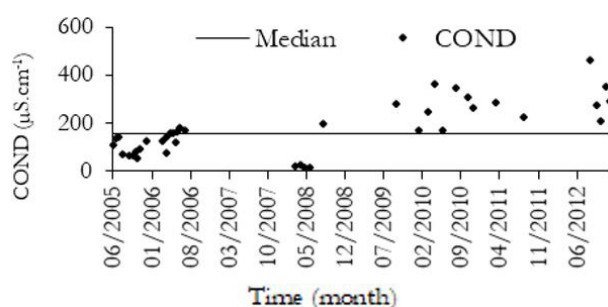
Figure 5. Median and conductivity time series at site IG4.

The analysis of historical series should be based on graphical analysis and knowledge about probable causes of non-homogeneity, for example, dates in which there were changes in methods, seasonality, among others.

An alternative to avoid the non-homogeneity of the COND series shown in Figure 5 would be to use part of the series, for example, from the year 2009, unless the objective is trend analysis.

Independence

Rejection of H_0 (independence) in favor of H_1 (dependence) was observed in 51% of the series. The test statistics can be seen in Table 5 and Table 6. Values well above the significance level (1.96) are observed, with emphasis on DBO, NH_4^+, N_{ORG} and COD, which are representative of the organic content of pollution.

For the tested series, it is observed that the irregular monitoring frequency in the longer series, predominantly with monthly and quarterly frequencies and long periods of time without data, results in high values of the test statistic with a maximum of 53.72 (Table 5), indicating a strong dependence relationship.

The results of the shorter series indicate that the monthly frequency also provides sufficient evidence to reject the hypothesis of independence in several series, however, with considerably lower values of the test statistic, with a maximum value of 17.18 (Table 6).

In water quality, when the sampling frequency is higher (daily, hourly, etc.), there is no evidence that the water quality will be better or worse in the next interval, since the concentrations are not only functions of the flows, but also of the highly variable contributions of loads in the system. On the other hand, when the frequency of monitoring is lower (monthly and quarterly), a certain behavior is expected in terms of water quality due to the seasonality of the climate and the hydrological cycles, which interfere with temperatures, rainfall volume and increased diffuse pollution in the rainier seasons.

These results indicate that the complexity in establishing the dependence ratio may be associated with the increase in the frequency of water quality monitoring. However, due to the seasonal effect, the longer series (greater than one year) may still be correlated. These relationships should be the object of future research.

An alternative for non-rejection of the independence hypothesis in the water quality time series is the analysis of data separated by season or, according to Gilbert (1987) and McBride (2005), the use of methods adapted to this characteristic or the use of techniques for the removal of seasonality and trends.

Table 4. Rejections of H_0 (homogeneity), in homogeneity test.

Variables	IG1	IG2A	IG2B	IG3	IG4	IG5	IG6
DQO	*	*	*	0.031	*	*	*
	*	*	0.016	0.034	*	*	*
DBO	0.002	0.017	*	0.026	0.017	*	0.035
	0.008	0.031	*	*	*	*	*
STT	*	0.022	*	*	*	*	*
	0.024	*	*	*	*	*	*
STV	*	*	*	*	*	*	*
	0.004	*	*	*	*	*	*
SDT	*	*	*	*	*	*	*
	0.020	*	*	*	*	*	*
SDV	*	*	*	*	*	*	*
	0.006	*	*	*	*	*	*
SST	*	0.000	*	0.000	*	*	*
	0.039	0.015	*	0.006	*	*	*
SSF	*	0.002	*	0.006	*	*	*
	*	0.049	*	0.029	*	*	*
SSV	*	0.005	*	0.019	*	*	*
	0.028	0.049	*	0.006	*	*	*
S_{SED}	*	*	*	0.013	*	*	*
	*	*	*	0.050	*	*	*
NH_4^+	0.042	*	*	*	*	*	*
	0.041	*	0.010	*	*	*	*
$N_{KEJLDAHL}$	*	*	*	*	*	*	*
	*	*	0.037	*	*	*	0.006
NO_2^-	0.000	0.050	*	0.000	0.000	0.002	0.006
	0.003	0.028	*	0.044	0.038	0.020	*
N_T	*	*	*	*	*	*	*
	0.034	*	0.037	*	*	*	*
P_T	0.000	*	*	*	0.027	*	0.033
	*	*	*	0.000	0.019	0.001	0.000
PO_4^{3-}	0.021	*	0.037	*	*	*	*
	*	*	0.006	*	*	*	*
P_{DISS}	*	*	*	*	*	*	*
	*	0.027	0.004	*	*	*	*
P_{PART}	*	*	*	0.043	*	0.012	0.008
	*	*	*	*	*	*	*
$P_{ORG\ DISS}$	*	*	*	*	*	*	0.021
	*	0.001	*	0.009	0.006	*	0.013
$COLIF_{FECAL}$	*	0.009	0.045	0.002	0.010	0.014	*
COD	0.021	0.008	*	*	*	0.020	0.004
	*	0.001	0.004	*	*	0.015	0.001
TURB	*	0.000	*	0.000	0.008	0.001	0.001
COND	**0.000**	**0.000**	*	**0.000**	**0.000**	**0.000**	**0.000**
OD	0.021	*	*	*	*	*	0.006
	*	*	*	*	*	*	*
$OD_{WINKLER}$	*	*	*	*	0.042	0.027	*
	*	*	*	0.025	*	0.005	0.005
pH	0.000	0.015	*	0.020	0.008	0.002	0.003
Q	0.001	*	0.037	*	*	*	*

* Non-rejection of H_0.

Stationarity

In Table 7 (sites with longer time series), H_0 (stationarity) rejections in favor of H_1 (trend) were observed in 31% of the series (420 series tested). Overall, higher values of the S statistic are associated with p-values well below 0.05 (Table 8). Of the total observed rejections, 55% indicate decreasing trend (negative) and 45% increasing trend (positive).

Table 5. Rejections of H_0 (independence), sites with longer time series.

Variable	IG1	IG2A	IG2B	IG3	IG4	IG5	IG6
DQO	5.71	3.26	4.71	3.28	**10.36**	5.74	2.15
	2.35	*	6.69	2.85	4.64	*	*
DBO	4.11	5.38	*	**12.48**	3.05	**12.63**	**17.70**
	*	2.57	*	**9.37**	*	**8.93**	3.95
STT	*	*	*	4.49	8.41		
	*	2.89	2.24	2.84	4.08	*	*
STF	9.89	2.45	3.04	2.17	6.25	*	4.06
	3.33	2.94	3.49	3.36	4.58	*	3.04
STV	*	5.52	5.78	7.77	*	9.54	*
	*	2.13	2.87	*	2.08	*	*
SDT	*	*	*	8.09	2.70	*	*
	*	3.23	*	2.48	2.88	*	*
SDF	10.73	3.31	*	2.38	2.30	3.11	3.76
	3.94	3.39	*	3.71	3.64	*	3.97
SDV	*	5.06	3.50	7.54	*	9.48	*
	*	*	*	*	*	*	*
SST	*	2.03	*	2.18	2.89	*	*
	*	2.50	7.82	2.24	2.28	*	*
SSF	*	*	3.62	*	2.62	*	*
	*	*	*	*	*	*	*
SSV	*	*	*	*	*	*	3.42
	2.11	3.51	5.98	3.30	3.12	*	2.24
S_{SED}	*	2.80	*	7.84	*	*	*
	*	*	3.63	*	*	4.53	*
NH_4^+	**11.44**	**12.28**	*	7.66	**10.94**	*	*
	5.19	6.22	3.81	5.97	**23.66**	7.32	7.90
N_{ORG}	**15.03**	**25.40**	*	**48.48**	**33.65**	**53.72**	**15.51**
	12.69	2.07	*	**20.03**	**18.59**	**39.25**	**23.56**
$N_{KEJLDAHL}$	7.08	5.27	*	4.66	*	5.05	3.42
	4.15	4.21	*	5.99	*	*	6.19
NO_2^-	3.39	6.02	4.72	5.33	6.29	*	3.74
	*	3.41	*	4.26	*	*	2.19
NO_3^-	*	*	*	2.17	*	4.11	**15.93**
	*	2.49	*	3.29	2.29	3.22	**10.42**
N_T	2.88	*	*	*	*	6.03	5.65
	2.01	*	*	*	*	*	*
P_T	*	3.94	*	*	*	*	2.09
	3.19	3.81	*	2.19	*	*	2.72
PO_4^{3-}	7.47	6.54	2.54	2.80	*	*	5.21
	*	5.35	3.13	*	*	2.71	6.19
P_{DISS}	3.35	9.32	2.08	*	2.13	*	9.79
	*	8.44	3.67	*	*	3.46	7.66
P_{PART}	2.84	*	*	*	2.26	*	*
	2.32	*	*	*	*	2.35	*
$P_{ORG\ DISS}$	2.04	8.87	*	3.63	2.46	2.72	7.38
	*	5.87	*	2.93	*	2.76	4.07
$COLIF_{TOTAL}$	*	*	6.19	3.25	3.23	4.09	*
$COLIF_{FECAL}$	*	*	*	2.25	3.28	9.05	3.95
COD	**15.16**	7.13	6.13	4.55	6.87	**12.53**	**13.13**
	7.24	*	*	2.54	**11.36**	**15.01**	*
TURB	3.64	*	3.20	2.12	2.35	2.26	*
COND	*	*	*	*	*	2.54	*
OD	*	*	2.98	*	*	2.11	4.57
	*	*	5.36	*	*	*	3.30
$OD_{WINKLER}$	*	*	*	2.92	*	2.22	*
	2.76	2.21	*	*	*	*	*
T	*	*	4.32	2.09	*	*	12.47
SECCHI	5.20	3.91	*	*	*	*	*
pH	*	*	*	**14.69**	3.91	4.91	7.43
Q	2.87	2.66	*	3.47	*	*	*

* Non-rejection of H_0.

Table 6. Rejections of H_0 (independence), sites with shorter time series.

Variables	IG7	PA	AT	BE	BA
DQO	5.36	*	3.54	*	5.21
	2.09	*	*	*	*
DBO	3.67	*	*	*	*
	5.51	*	*	*	*
STT	4.67	3.01	*	*	*
	*	3.06	*	*	*
STF	3.73	2.88	*	*	*
	2.27	3.11	17.18	*	*
STV	2.49	*	*	2.91	2.41
	*	2.13	3.38	*	*
SDT	3.15	*	*	*	*
	*	*	*	*	*
SDF	*	*	*	*	*
	2.27	*	*	*	*
SDV	2.30	*	*	2.06	*
	*	*	*	*	*
SST	*	3.84	*	2.12	5.12
	*	3.80	2.17	1.96	*
SSF	*	4.19	1.99	*	3.87
	*	4.07	1.97	*	*
SSV	*	*	*	2.34	8.85
	*	*	3.18	2.00	*
S_{SED}	*	6.04	*	*	*
	*	5.62	*	*	*
NH_4^+	2.26	2.26	2.53	*	2.80
	2.70	*	*	*	*
$N_{KEJLDAHL}$	*	*	*	*	*
	2.11	*	*	*	*
NO_2^-	*	6.41	2.62	*	2.79
	*	8.77	2.86	*	*
N_T	2.08	*	*	*	*
	*	*	*	*	*
PO_4^{3-}	2.00	*	*	*	*
	*	*	*	*	*
P_{DISS}	*	*	*	1.99	*
P_{PART}	*	5.40	*	2.51	4.22
	*	4.33	*	2.32	*
$COLIF_{TOTAL}$	3.46	*	*	*	14.44
$COLIF_{FECAL}$	*	5.12	14.35	2.10	*
COD	2.33	*	10.15	*	5.23
	7.05	4.88	2.64	2.51	*
TURB	3.86	2.29	4.37	*	*
COND	*	3.49	5.93	*	*
OD	*	2.39	*	*	*
	2.01	*	*	1.99	*
$OD_{WINKLER}$	3.37	*	*	*	*
	*	*	*	*	3.31
T	3.29	2.44	2.24	3.33	2.67
pH	1.97	*	*	4.96	2.69

* Non-rejection of H_0.

In Table 9 (sites with shorter time series), there is an increasing trend in some sites and variables as the data approaches the summer (dates between 08/12 and 12/12), but with values of S associated with p-values close to 0.05. The term trend, in essence, refers to long-term changes, and caution should be taken in this interpretation.

Table 7. Rejections of H_0 (stationarity), sites with longer time series, S statistic.

Variables	IG1	IG2A	IG2B	IG3	IG4	IG5	IG6
DQO	*	*	*	247	*	*	*
	-60	-144	-28	*	*	*	*
DBO	*	-156	*	-162	-170	-207	-162
	-113	-250	*	-124	*	*	*
STT	*	283	*	210	*	*	*
	-56	*	*	*	*	*	*
STF	*	219	*	227	*	169	*
	*	*	*	*	*	*	*
STV	-147	*	*	*	*	*	*
	-70	*	*	*	*	*	*
SDV	-151	*	*	*	*	*	*
	-79	*	*	*	*	*	*
SST	*	379	*	364	*	*	*
	-61	*	*	*	*	*	*
SSF	*	309	*	309	*	*	*
	*	*	*	117	*	*	*
SSV	*	226	*	206	*	*	*
	*	*	*	*	*	*	*
S_{SED}	*	*	*	93	*	*	26
	*	*	*	68	*	*	*
NH_4^+	-218	-170	-32	-151	-175	-181	*
	-68	-190	-50	*	*	-179	*
N_{ORG}	*	*	*	155	141	140	165
	*	*	*	*	*	*	116
$N_{KEJLDAHL}$	*	*	*	*	*	*	*
	*	-136	*	*	*	*	120
NO_2^-	-450	-282	*	-422	*	-388	-260
	-115	-279	*	-209	206	-305	-205
NO_3^-	*	*	*	-138	*	*	*
	*	*	*	*	-92	*	*
N_T	71	*	*	*	*	59	41
	*	*	*	*	*	*	*
P_T	299	196	*	285	266	225	256
	*	*	-32	269	122	245	257
PO_4^{3-}	-35	*	-26	*	*	*	*
	*	*	-34	*	*	*	*
P_{DISS}	*	*	*	*	*	*	*
	*	-62	-40	*	*	*	-41
P_{PART}	*	*	*	46	*	55	35
	*	*	*	*	*	*	*
$P_{ORG\ DISS}$	*	*	*	*	*	*	*
	*	-62	*	-33	-23	*	-49
$COLIF_{FECAL}$	*	53	27	63	61	54	23
COD	*	-119	*	*	*	-74	-96
	-81	-122	-42	*	*	-79	-82
TURB	*	293	*	*	175	213	219
COND	252	404	*	*	389	405	376
$OD_{WINKLER}$	*	*	*	-34	*	-59	-58
	*	-41	*	-32	*	-61	-55
SECCHI	*	-73	*	-104	*	*	*
pH	*	314	*	290	289	300	245
Q	*	-186	-32	*	*	*	*

* Non-rejection of H_0.

Table 8. Rejections of H_0 (stationarity), sites with longer time series, p-values.

Variables	IG1	IG2A	IG2B	IG3	IG4	IG5	IG6
DQO	*	*	*	0.002	*	*	*
	0.037	0.031	0.032	*	*	*	*
DBO	*	0.035	*	0.026	0.020	0.006	0.011
	0.001	0.001	*	0.028	*	*	*
STT	*	0.001	*	0.009	*	*	*
	0.49	*	*	*	*	*	*
STF	*	0.007	*	0.006	*	0.025	*
	*	*	*	*	*	*	*
STV	0.044	*	*	*	*	*	*
	0.018	*	*	*	*	*	*
SDV	0.035	*	*	*	*	*	*
	0.009	*	*	*	*	*	*
SST	*	0.000	*	0.000	*	*	*
	0.046	*	*	*	*	*	*
SSF	*	0.000	*	0.000	*	*	*
	*	*	*	0.050	*	*	*
SSV	*	0.006	*	0.011	*	*	*
	*	*	*	*	*	*	*
S_{SED}	*	*	*	0.025	*	*	0.040
	*	*	*	0.049	*	*	*
NH_4^+	0.006	0.020	0.016	0.040	0.018	0.015	*
	0.022	0.005	0.000	*	*	0.013	*
N_{ORG}	*	*	*	0.036	0.039	0.046	0.010
	*	*	*	*	*	*	0.037
$N_{KEJLDAHL}$	*	*	*	*	*	*	*
	*	0.039	*	*	*	*	0.033
NO_2^-	0.000	0.001	*	0.000	*	0.000	0.000
	0.000	0.000	*	0.002	0.000	0.000	0.002
NO_3^-	*	*	*	0.037	*	*	*
	*	*	*	*	0.036	*	*
N_T	0.002	*	*	*	*	0.008	0.036
	*	*	*	*	*	*	*
P_T	0.000	0.014	*	0.001	0.001	0.005	0.000
	*	*	0.016	0.000	0.025	0.001	0.000
PO_4^{3-}	0.037	*	0.043	*	*	*	*
	*	*	0.012	*	*	*	*
P_{DISS}	*	*	*	*	*	*	*
	*	0.005	0.004	*	*	*	0.024
P_{PART}	*	*	*	0.032	*	0.013	0.046
	*	*	*	*	*	*	*
$P_{ORG\ DISS}$	*	*	*	*	*	*	*
	*	0.002	*	0.040	0.043	*	0.003
$COLIF_{FECAL}$	*	0.002	0.020	0.000	0.001	0.002	0.011
COD	0.012	0.000	*	*	*	0.020	0.002
	*	0.000	0.002	*	*	0.014	0.004
TURB	*	0.000	*	*	0.017	0.005	0.001
COND	0.001	0.000	*	*	0.000	0.000	0.000
$OD_{WINKLER}$	*	*	*	0.034	*	0.002	0.001
	*	0.023	*	0.016	*	0.001	0.001
SECCHI	*	0.044	0.045	*	*	*	*
pH	*	0.000	*	0.001	0.000	0.000	0.001
Q	*	0.010	0.016	*	*	*	*

* Non-rejection of H_0.

From IG1 to IG6, increasing trends are concentrated on variables such as N_{ORG}, N_T, P_T, $COLIF_{FECAL}$, COND, TURB and pH, indicating the increasing contribution of organic matter in the basin. Decreasing trends are observed in variables such as DBO, NH_4^+, NO_2^-, PO_4^{3-}, P_{DISS}, $P_{ORG\ DISS}$ and COD, which represent the products of decomposition.

The increasing contribution associated with the reduction of decomposition products indicates a possible reduction of

Table 9. Rejections of H_0 (stationarity), sites with shorter time series, S statistic (p-value).

Variable	IG7	PA	AT	BE	BA
STT	* *	* 8 (0.043)	* *	8 (0.042) 8 (0.042)	* 8 (0.042)
STF	* *	* 8 (0.043)	* *	* 8 (0.042)	* *
STV	* *	8 (0.042) 10 (0.014)	* *	* *	* *
SDT	* *	8 (0.042) 8 (0.042)	* *	* 10 (0.008)	* 8 (0.042)
SDF	* *	* *	* 10 (0.008)	* 8 (0.042)	* 8 (0.042)
SDV	* *	8 (0.042) 8 (0.042)	* *	* *	* *
N_{ORG}	* *	* *	* 8 (0.042)	* *	* 8 (0.042)
NO_2^-	* *	* 8 (0.043)	* *	* *	* *
N_T	* *	* *	* *	* 8 (0.042)	* 8 (0.042)
P_T	* *	* 10 (0.008)	* *	* 8 (0.042)	* *
P_{PART}	* *	* *	* *	* *	-8 (0.043) *
COD	* *	-6 (0.045) *	* *	* 10 (0.014)	* 8 (0.042)
TURB	*	*	*	*	-8 (0.042)
COND	*	8 (0.043)	6	*	*
T	10 (0.014)	10 (0.014)	8 (0.043)	*	8 (0.043)

* Non-rejection of H_0.

Table 10. Comparison among rejections in tests for flow Q.

Test	IG1	IG2A	IG2B	IG3	IG4	IG5	IG6
R	0.000	0.000	*	0.040	*	*	*
H	0.001	*	0.037	*	*	*	*
I	2.87	2.66	*	3.47	*	*	*
T	*	-186	-32	*	*	*	*
S	KM	*	*	*	*	*	*

R: randomness, H: homogeneity, I: independence, T: trend, S: seasonality, KM: rejection of hypothesis of equality among seasons by Kruskal-Wallis and median tests. * Non-rejection of H_0.

biodegradability, consistent with the increasing trend observed in the pH variable, which negatively affects the action of microorganisms.

Synthesis of results

Almost all water quality variables were rejected in at least one of the assumptions at most monitoring sites. Some series, such as the STT and Q at IG6, met all the assumptions and did not reflect seasonal effects. It is important to note that these series were obtained by the same monitoring strategy and under the same seasonal conditions as the other series, evidencing the existence of other factors that determine their attendance. The absence of seasonality is associated with the non-representativeness, mainly, of the flow series. The flow series in the Iguaçu River basin show seasonal variations.

The understanding of the applied methods allowed the establishment of relations of cause and effect between randomness, homogeneity and independence and presence or not of seasonality and trends. Seasonality may give rise to cyclical (non-random) variability, to data significantly higher or lower in different (heterogeneous) seasons, and to data decreasing or increasing in a chronological sequence as they approach a particular season of the year (dependence).

The presence of trends may produce significant differences between recent and old data (non-homogeneity), its slope may result in few runs (not randomness), and make predictable part of the variability of future data depending on the last data obtained (dependence).

However, the results show that seasonality and trend can not assess the assumptions of randomness, homogeneity, and independence alone. Table 10 shows that the absence of seasonality and trends in the flow series are associated with compliance with all assumptions in IG4, IG5 and IG6, but not in IG3. The lack of trend may be associated to non-attendance, as in IG1, but also to compliance, as in IG4, IG5 and IG6. In this way, it is understood that all assumptions must be tested prior to frequency analysis in water quality time series. Some causes of non-compliance with basic assumptions can be verified graphically as well as possible solutions to work around the problem.

CONCLUSIONS

The verification of the basic assumptions for frequency analysis in water quality series in the Upper Iguaçu Basin evidenced the need for a more cautious approach to the analysis of large databases. The assumption of randomness, homogeneity, independence and stationarity premises, typical in studies of water quality, can be misleading in most of the series.

The applications of the tests allowed the identification of series important characteristics, as well as, reveled deficiencies of the typical strategies of water quality monitoring. The fulfillment of the basic assumptions for the analysis of water quality time series is essentially related to the monitoring strategies and techniques of approach of the series for analysis. As surface water quality variability is strongly related to anthropogenic activities, monitoring strategies should generate information on the cycles of variation corresponding to the various activities of man, i.e. daily and weekly cycles, as well as holidays.

Likewise, the annual cycle, due to seasonality, and other events that significantly interfere with water quality conditions through diffuse pollution should also be monitored. The data for each cycle or event should be compared and, if they are significantly different, separately analyzed.

However, due to the high costs, the intrinsic difficulties of monitoring and the lack of planning, the typical monitoring

strategies involve irregular frequencies, monitoring only on working days (representing around 24% of the time of one year), absence of collection in periods of heavy rain, night periods, holidays, weekends and summer, which together represent most of the time in a year.

The time series generated from these strategies represent only part of the daily cycle (commercial hours) and their seasonality, not being able to provide information on other periods. Direct analysis of these series often produces non-representative statistics, as they do not meet the basic assumptions of randomness, homogeneity, independence and stationarity.

Among the possible solutions are the separation of the data into groups, the application of trend and seasonality removal techniques and the transformation into annual series by means of averages or medians. However, irregular frequencies, without planning, negatively interfere in this process. In general, there is little data in the summer and data are concentrated irregularly between stations. This makes it difficult to characterize the seasonality and makes the mean and median biased.

The results point to the need to rethink the way water quality monitoring networks operate, especially regarding the frequency of sample collection, in order to have representative data.

The knowledge on this topic in the area of water quality is still incipient, with many questions still unanswered, such as: which tests are most appropriate for checking the assumptions? how do the sampling frequencies interfere in the tests of randomness, homogeneity, independence and stationarity? what other factors besides seasonality and trends lead to non-randomness, non-homogeneity, dependence and non-stationarity? what is the best strategy for monitoring and analyzing data when resources are scarce?

ACKNOWLEDGEMENTS

REFERENCES

AMERICAN PUBLIC HEALTH ASSOCIATION – APHA. American Water Works Association – AWWA. Water Environment Federation – WPCF, *Standard methods for the examination of water and wastewater*. Washington, 1999.

CHADEE, J. C.; SHARMA, C. Fitting flexible wind speed models: a new approach. In: WELLS, J. (Ed.). *Solutions for energy security & facility management challenges*. Lilburn: The Fairmont Press, 2003. p. 249.

COELHO, M. *Estratégia de monitoramento da qualidade da água para a gestão de recursos hídricos em bacias urbanas*. 2013163 f. Dissertation (Environmental and Water Resources Engineer) - Federal University of Paraná, Curitiba, 2013.

COSTA, M. P.; OLIVEIRA, R. B. S.; SOUZA, M. L. Análise da Tendência do Índice de Qualidade das Águas da Região Hidrográfica do Paraná no Período 2000-2009. In: BRAZILIAN SYMPOSIUM ON WATER RESOURCES, 29., 2011, Maceió. *Proceedings...* São Paulo: ABRH, 2011. Available from: <http://www.abrh.org.br/SGCv3/index.php?PUB=3&ID=81&PAG=1>. Access on: 26 may 2015.

DAHM, K. G.; GUERRA, K. L.; MUNAKATA-MARR, J.; DREWES, J. E. E. *Trends* in water quality variability for coalbed methane produced water. *Journal of Cleaner Production*, v. 84, p. 840-848, 2014. http://dx.doi.org/10.1016/j.jclepro.2014.04.033.

FORMIGONI, Y., BRITES, A. P. Z.; FERNANDES, C. V. S.; PORTO, M. F. A. Análise crítica da curva de permanência de qualidade da água com base em dados históricos. In: BRAZILIAN SYMPOSIUM ON WATER RESOURCES, 29., 2011, Maceió. *Proceedings...* São Paulo: ABRH, 2011. Available from: <http://www.abrh.org.br/SGCv3/UserFiles/Sumarios/ec99f1061725d63695b7bea3067296d4_31aef3f050dfe0930e76fff29fa8b397.pdf>. Access on: 26 may 2015.

GILBERT, R. O. *Statistical methods for environmental pollution monitoring*. John Wiley & Sons, 1987.

GONÇALVES, V. D., FERNANDES, C. V. S.; KNAPIK, H. G.; MÄNNICH, M. Desafios da gestão de recursos hídricos: um olhar sobre o monitoramento ambiental de rios. In: BRAZILIAN CONGRESS OF SANITARY AND ENVIRONMENTAL ENGINEERING, 26., 2011, Porto Alegre. *Proceedings...* São Paulo: ABES, 2011.

HAMED, K. H.; RAO, A. R. A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology*, v. 204, n. 1-4, p. 182-196, 1998. http://dx.doi.org/10.1016/S0022-1694(97)00125-X.

HELSEL, D. R.; HIRSCH, R. M. *Statistical methods in water resources*: techniques of water resources investigations of the United States Geological Survey. Reston: US Geological Survey, 2002. chap.**A3**. Book 4: Hydrologic analysis and interpretation. Available from: <http://pubs.usgs.gov/twri/twri4/>. Access on: 2 apr. 2013.

HIRSCH, R. M.; MOYER, D. L.; ARCHFIELD, S. A. Weighted Regressions on Time, Discharge, and Season (WRTDS), with an Application to Chesapeake Bay River Inputs1. *JAWRA Journal of the American Water Resources Association*, v. 46, n. 5, p. 857-880, 2010. PMid:22457569. http://dx.doi.org/10.1111/j.1752-1688.2010.00482.x.

HULLEY, M.; CLARKE, C.; WATT, E. Low flow frequency analysis for streams with mixed populations. *Canadian Journal of Civil Engineering*, v. 42, n. 8, p. 503-509, 2015. http://dx.doi.org/10.1139/cjce-2014-0323.

KNAPIK, H. G.; FERNANDES, C. V. S.; AZEVEDO, J. C. R. Caracterização e monitoramento de matéria orgânica em rios: aplicabilidade na gestão de recursos hídricos. In: BRAZILIAN SYMPOSIUM ON WATER RESOURCES, 20.. 2013, Bento Gonçalves. *Proceedings...* São Paulo: ABRH, 2013. Available from:

<http://www.abrh.org.br/SGCv3/index.php?PUB=3&ID=1 55&PUBLICACAO=SIMPOSIOS>. Access on: 26 may 2015.

KONDAGESKI, J. H.; FERNANDES, C. V. S. Calibração de Modelo de Qualidade da Água para rio utilizando Algoritmo Genético: Estudo de caso de Rio Palmital-PR. *Brazilian Journal of Water Resources*, v. 14, n. 1, p. 63-73, 2009.

MARIN, M. C. F C.; SCUISSIATO, C.; FERNANDES, C. V. S.; PORTO, M. F. A. Proposta preliminar de reenquadramento dos corpos de água em classes e avaliação do seu risco de não atendimento: estudo de caso da Bacia do Alto Iguaçu. In: BRAZILIAN SYMPOSIUM ON WATER RESOURCES, 17., 2007, São Paulo. *Proceedings...* São Paulo: ABRH, 2007. Available from: http://www.abrh.org.br/SGCv3/index.php?PUB=3&ID= 155&PUBLICACAO=SIMPOSIOS>. Access on: may 26, 2015.

MCBRIDE, G. B. *Using statistical methods for water quality management*. New Jersey: John Wiley & Sons, 2005.

MERZ, B.; THIEKEN, A. H. Separating natural and epistemic uncertainty in flood frequency analysis. *Journal of Hydrology*, v. 309, n. 1-4, p. 114-132, 2005. http://dx.doi.org/10.1016/j.jhydrol.2004.11.015.

NAGHETTINI, M.; PINTO, E. J. A. *Hidrologia e estatística. geological survey of Brazil*. Belo Horizonte: CPRM, 2007.

PITRAT, D. M. J. J. *Avaliação da contaminação por metais em rios: estudo de caso da Bacia do Rio Passaúna*. 2010. 231 f. Dissertation (Water Resources and Environmental Engineering) - Federal University of Paraná, Curitiba, 2010.

SIEGEL, S.; CASTELLAN JUNIOR, N. J. *Estatística não paramétrica para ciências do comportamento*. 2nd ed. Porto Alegre: Artmed, 2006.

TORRES, I. C.; LEMOS, R. S.; MAGALHÃES JUNIOR, A. P. Influence of the Rio Taquaraçu in the water quality of the Rio das Velhas: subsidies for reflections of the case of water shortage in Belo Horizonte metropolitan region–MG, Brazil. *Brazilian Journal of Water Resources*, v. 21, n. 2, p. 429-438, 2016.

VIEIRA, J.; FONSECA, A.; VILAR, V. J. P.; BOAVENTURA, R. A. R.; BOTELHO, C. M. S. Water quality modelling of Lis River, Portugal. *Environmental Science and Pollution Research International*, v. 20, n. 1, p. 508-524, 2013. PMid:23001788. http://dx.doi.org/10.1007/s11356-012-1124-5.

WANG, Y.; WANG, P.; BAI, Y.; TIAN, Z.; LI, J.; SHAO, X.; MUSTAVICH, L. F.; LI, B.-L. Assessment of surface water quality via multivariate statistical techniques: a case study of the Songhua River Harbin region, China. *Journal of Hydro-environment Research*, v. 7, n. 1, p. 30-40, 2013. http://dx.doi.org/10.1016/j.jher.2012.10.003.

Authors contributions

Marcelo Coelho: Literature review, writing, running the statistical tests, analysis of results, method development, conclusions and translation.

Cristovão Vicente Scapulatempo Fernandes: Advisor, analysis of results and conclusions and translation.

Daniel Henrique Marco Detzel: Advisor, analysis of results and conclusions and translation.

Michael Mannich: Structure of the paper, analysis of results and conclusions.